Differences in local population history at the finest level: the case of the Estonian population

Vasili Pankratov, Francesco Montinaro, Alena Kushniarevich, Georgi Hudjashov, Flora Jay, Lauri Saag, Rodrigo Flores, Davide Marnetto, Marten Seppel, Mart Kals, et al.

European Journal of Human Genetics. Published 25 July 2020. https://doi.org/10.1038/s41431-020-0699-4

Title: Differences in local population history at the finest level: the case of the Estonian population

Running title: Genetic structure of Estonia

Vasili Pankratov[a,], Francesco Montinaro[a], Alena Kushniarevich[a], Georgi Hudjashov[a,b], Flora Jay[c], Lauri Saag[a], Rodrigo Flores[a], Davide Marnetto[a], Marten Seppel[d], Mart Kals[a], Urmo Võsa[a], Cristian Taccioli[e], Märt Möls[f], Lili Milani[a], Anto Aasa[g], Daniel John Lawson[h], Tõnu Esko[a], Reedik Mägi[a], Luca Pagani[a,e,1], Andres Metspalu[a,1], Mait Metspalu[a,1]*

[a]Institute of Genomics, University of Tartu, Tartu, 51010, Estonia;

[b]Statistics and Bioinformatics Group, School of Fundamental Sciences, Massey University, Palmerston North 4474, New Zealand;

[c]Laboratoire de Recherche en Informatique, CNRS UMR 8623, Université Paris-Sud, Université Paris-Saclay, Orsay 91405, France

[d]Institute of History and Archaeology, University of Tartu, Tartu 51005, Estonia;

[e] Department of Biology, University of Padova, Padova 35131, Italy;

[f]Institute of Mathematical Statistics, University of Tartu, Tartu 50409, Estonia;

[g]Institute of Geography University of Tartu, Tartu 51003, Estonia;

[h]Medical Research Council Integrative Epidemiology Unit, Department of Population Health Sciences, Bristol Medical School, University of Bristol, Bristol BS8 2BN, United Kingdom;

*Corresponding author vasilipankratov@gmail.com

[1]Contributed equally

European Journal of Human Genetics. Published 25 July 2020. https://doi.org/10.1038/s41431-020-0699-4

Abstract

Several recent studies detected fine-scale genetic structure in human populations. Hence, groups conventionally treated as single populations harbour significant variation in terms of allele frequencies and patterns of haplotype sharing. It has been shown that these findings should be considered when performing studies of genetic associations and natural selection, especially when dealing with polygenic phenotypes. However, there is little understanding of the practical effects of such genetic structure on demography reconstructions and selection scans when focusing on recent population history. Here we tested the impact of population structure on such inferences using high-coverage (~30X) genome sequences of 2,305 Estonians. We show that different regions of Estonia differ in both effective population size dynamics and signatures of natural selection. By analyzing identity-by-descent segments we also reveal that some Estonian regions exhibit evidence of a bottleneck 10-15 generations ago reflecting sequential episodes of wars, plague, and famine, although this signal is virtually undetected when treating Estonia as a single population. Besides that, we provide a framework for relating effective population size estimated from genetic data to actual census size and validate it on the Estonian population. Our results suggest that the history of human populations within the last few millennia can be highly region-specific and cannot be properly studied without taking local genetic structure into account. Our approach to estimating the census population size may be widely used both to cross-check estimates based on historical sources as well as to get insight into times and/or regions with no other information available.

Main text

As more and more datasets including genetic data from hundreds and thousands of individuals become available it becomes apparent that most if not all human populations exhibit at least some degree of geography-driven genetic structure even at small scales (for some examples see[1–5]). Many recent publications have shown the confounding effect of such population structure on studies of genetic associations and natural selection signals, mainly in the case of polygenic phenotypes[6–9]. Here we study the fine-scale genetic structure of the Estonian population and the local differences in recent demographic history and action of natural selection between genetically defined Estonian subgroups to gain a deeper understanding of the forces shaping this population structure and the consequences it has for population genetics analyses. In doing so we make use of high-coverage whole genome sequences from more than 2300 Estonian individuals generated within a different study[10].

Our exploratory principal component analysis (PCA) (Figure 1) shows the presence of genetic structure within Estonia with the main differentiation between South-East and North-East of the country in agreement with previous studies[2,11,12]. To zoom-in into the fine-scale structure in Estonia, we used total genetic length of shared IBD segments detected with *IBDseq*[13] as input for the fineSTRUCTURE[14] clustering algorithm (Methods). We applied this approach to a subset of 468 individuals sampled in rural areas at the age of 50 or more, as this cohort is expected to be the least affected by recent migrations (Figure 2, SI1:2.3). We refer to this subset as “R50+” throughout the text (Methods).

IBD-based analysis (Figure 2) reinforces previous observations[2,11,12], including the strong differentiation between South-East and the rest of Estonia, and provides a deeper insight into Estonian genetic structure, showing that most of the revealed clusters are highly geographically localized. The sharing matrix provides additional details. First, off-diagonal sharing also reflects geography with clusters from the same area tending to have higher inter-cluster sharing. Second, intra-cluster sharing substantially varies among clusters,

implying differences in effective population size (Ne), which is also supported by the results of homozygosity-by-descent analysis (Figure 3).

In order to understand how gene flow barriers and/or differences in local population density shaped the IBD-sharing pattern in the R50+ dataset, we inferred migration surfaces using MAPS[15]. We used two windows of IBD segments length (in centimorgans, cM), 2-6 cM and more than 6 cM, which under a simplistic model of infinite population size have mean segment ages of 50 and 12.5 generations respectively[15]. Results for the two length bins generally agree with each other, suggesting higher levels of gene flow in the North along with a barrier separating South-East Estonia (SI1:2.4). A second barrier, separating the islands, especially Hiiumaa, from the mainland is also evident. This observation suggests that the population ancestral to modern South-East Estonians was partially isolated from the rest of the country at least since 50 generations ago. Interestingly, this genetic differentiation is consistent with linguistic data suggesting that the deepest split within the Finnic languages separates Southern Estonian from the other branches of the phylum that includes Northern Estonian[16].

As local differences in admixture with external populations may have played a role in creating the observed genetic structure within Estonia we looked at patterns of haplotype sharing between R50+ Estonians and different non-Estonian populations (Table SI2:3.1-I). Here we used a conventional CHROMOPAINTER/fineSTRUCTURE/GLOBETROTTER (CP/FS/GT) approach[17] (Methods). Figure 4 shows the results of non-negative least squares (NNLS)[1], modelling each individual from the R50+ dataset as a result of admixture between non-Estonian groups revealed by CP/FS (Figure 4, SI1:3.1, Table SI2-3.1-IV).

Admixture signals in Figure 4 show clear geographic patterns that match known historical evidence of external migration to Estonia, including Swedish settlements on the western coast and islands in 14-15th centuries and Finnish immigration to North-East Estonia in the 17th century[18]. Comparing NNLS results between clusters from Figure 2 we found that some of them, such as NE_1 and NE_2, stand out in terms of sharing with external groups but most of the clusters have overlapping distributions of NNLS scores (SI1:3.1). A similar

pattern is observed in IBD-sharing patterns (SI1:3.2). These results suggest that admixture with non-Estonian groups can only partially explain the fine genetic structure observed in Figure 2.

We show that, despite the small territory it occupies, the Estonian population exhibits a readily detectable genetic structure, reflected in patterns of IBD segments sharing (Figure 2) and allele frequencies (Figure 1, Table SI2-2.3-III, Table SI2-2.3-IV). Next, we sought to explore whether this differentiation has any effect on the reconstruction of demographic processes, namely whether there are region-specific differences in effective population size dynamics and action of natural selection. We hence applied *IBDNe*, which estimates effective population size (Ne) in past generations[19], and SDS (Singleton Density Score), a tool for detecting signatures of natural selection[20], as both methods give insight into very recent time periods, when regional differences in population history may be anticipated. For both analyses, we used the entire dataset of 2,305 samples, for which clusters were inferred using the same approach as for the R50+ subset (Figure 5).

We ran *IBDNe*[19] on the four most distinct clusters from Figure 5, representing four regions of Estonia: North-West, North-East, South-West and South-East and observed rather distinct Ne trajectories (Figure 6a, SI1:4.2). In particular, all clusters (except for eSE_5) show evidence of an effective population size decline between 10 and 20 generations ago, which is not detected when the entire dataset is analyzed (Figure 6a). Overall, these results suggest that population dynamics are region-specific and hence population-wide result may depend on the sampling scheme. For a deeper understanding of this phenomenon and the effects of other factors on *IBDNe* results, we applied *IBDNe* to genetic data, simulated under various demographic scenarios (SI1:4.1). Furthermore, the same approach has been applied to genotype data from the UK population, where regional differences in Ne dynamics are observed as well (SI1:4.3).

Based on our simulations and MAPS results, we propose that most of the differences in Ne dynamics between Estonian subpopulations may be attributed to different patterns of gene flow and external admixture. South-West and North-West Estonia are characterized by an

overall high level of gene flow (SI1:2.4), leading to similar Ne trajectories that deviate only during the last 20 generations (Figure 6, SI1:4.2) reflecting very recent differences in population size dynamics and/or migration. This also brings about the idea that the strong bottleneck in South-West could contribute to the observed population structure, in particular leading to differentiation of South-West and its subgroups. On the other hand, South-East Estonia has the most distinct Ne trajectory according to Figure 6a, having a substantially lower long-term Ne compared to other regions. Together with MAPS results (SI1:2.4) this might suggest a recent expansion of a previously small-size eSE_5-like population and its admixture with other local subpopulations occupying South-East Estonia thus contributing to other eSE groups. This, in turn, results in a rather recent increase in relative proportion of individuals with eSE_5-like ancestry in the entire Estonian population affecting the Ne reconstructions for the entire dataset (SI1:4.2).

Given our understanding of confounders of the observed regional Ne patterns, we exploited the fine-grained temporal resolution enabled by *IBDNe* to infer changes in actual census sizes (Nc) of the ancestors of contemporary Estonians, adapting previous theoretical work[21] to empirical case of human populations (Methods). We applied equation [3] (Methods) to the Estonian-wide Ne trajectory inferred using the Est1527 subset, which excludes clusters that can be considered as outliers in terms of external admixture and/or Ne trajectory (SI1:4.4). We then compared the inferred Nc with available historical estimates (Figure 6b) showing a remarkable match between the two with the exception of the last three generations, for which *IBDNe* estimates are extrapolated from preceding time points[19]. This match may be attributed to i) our success in adequately controlling for events of recent gene flow and population structure; ii) the relatively recent time intervals considered, which limits the range of spatial interaction among the ancestors of contemporary Estonians. However, note that the pronounced fluctuations in Nc reported by historians between 1500 and 1700 are only very roughly approximated by the Ne-derived curve which, as expected[22], provides only relatively long-term harmonic average of Ne. Nevertheless, we suggest that after controlling for confounders such as population structure and admixture and keeping in mind all the

assumptions implied by the biological notion of Ne, our approach could be used to convert Ne to human Nc at any time interval for which historical records are missing, including the ones provided by PSMC[23], which are beyond the scope of the current paper.

We then questioned whether natural selection could have also acted differently within the Estonian population. In doing so, we applied singleton density score (SDS)[20] to the entire dataset of 2,305 samples as well as to two regional subsets, South-East Estonia (SE, consisting of 1,029 samples belonging to clusters eSE_1 - eSE_5 in Figure 7) and the remaining 1,276 samples from the rest of the country (nonSE) (Methods, SI1:5.1).

First, we inspected the genome-wide distribution of positive SDS scores in the three datasets (Figure 7) for any evidence of recent selection acting at individual loci.

Unlike other studies that used SDS[20,24] we don't observe any hits with very low p-value (possible reasons are discussed in SI1:5.3). However, we see that the distribution of SDS scores differs between the three datasets (Figure 7, Table SI2-5.3-I). Whereas one genome-wide significant hit (rs75386033 and rs79907158 on chromosome 6) is detected in the SE, nonSE and the entire dataset had many more hits with p-values in the range between $5x10^{-8}$ and $1x10^{-5}$ (Figure 7, SI1:5.3, SI1:5.4, Table SI2-5.3-I). Whereas most of the top SDS signals do not overlap between the three datasets analyzed, one region on chromosome 10 corresponding to the *WDFY4* gene appears in both SE and nonSE (Figure 7, Table SI2-5.3-I). It has been shown that *WDFY4* is involved in immune response toward viral and tumor antigens[25] as well as in autoimmune diseases[26–28]. Functional annotation of variants with positive SDS scores[29] coupled with enrichment test[30,31] did not reveal any annotation category to be specific for a particular subset studied (S1:5.3, S1:5.4, Tables SI2-5.3-IV-V, Tables SI2-5.4-IV-VII). Likewise, alternative enrichment test employing the GWAS catalog showed that similar phenotype categories are present in the three tested datasets (S1:5.3, S1:5.4, Table SI2:5.3-III, Tables SI2:5.4-II-III).

On the other hand, frequency differences of rs75386033 derived allele T (10.3% in SE vs 6.1% in nonSE, Weir and Cockerham[32] Fst=0.0117 corresponding to the 0.999 percentile of the genome-wide distribution (Figure S5.4-I) together with its low standardized SDS p-value

form strong evidence for a recent frequency increase of the rs75386033 T allele in South-East Estonia. Both rs75386033 and rs79907158 lie within an intron of the GRM1 gene, which is characterized by high levels of expression in the brain (https://www.ncbi.nlm.nih.gov/gene/2911). These SNPs themselves are not known to be associated with any phenotypes, however, there are some indications that variant rs362870 which is in high linkage disequilibrium with rs75386033 and rs79907158, might be a cis-eQTL for the *EPM2A* gene (SI1:5, Table SI2:5-II), suggesting a plausible biological effect behind the frequency change. *EPM2A* gene is associated with Lafora disease which is a form of progressive myoclonus epilepsy[33–35]. This gene codes for a protein called laforin which is involved in regulating glycogen synthesis and potentially prevents glycogen accumulation in neurons[33–35].

Given the lack of information on the phenotypic effect of this *GRM1* allele and its modestly strong SDS signal, it is unclear whether the raise in frequency happened due to actual selection or because of random genetic drift especially given the fact that South-East Estonians exhibit signals of long-lasting low Ne and further differentiation into smaller subclusters (Figures 2 and 6). Nevertheless, differential SDS signals between the entire Estonian dataset and its subsets including *GRM1*, *WDFY4*, suggest that recent selection, restricted to regional subpopulations, may remain undetected if population-wide datasets are treated as a single entity.

In conclusion, here we describe a dataset of more than 2300 high-coverage Estonian genomes from a population genetics perspective making it one of the smallest populations to date with such high-resolution data available. We show that the Estonian population, despite occupying a small area with no strong geographic barriers, is genetically structured and exhibits rather pronounced interregional differences with respect to recent admixture with neighbouring populations, population dynamics and potential action of natural selection. These observations together with results of other studies suggest that population stratification could be ubiquitous in human populations, and should be taken into account in any large-scale genetic study including reconstructions of recent population history. We also

show that we are able to accurately link effective population size to actual census size based on some simple assumptions about human population age structure and reproduction patterns. We envisage future studies exploiting this framework to ultimately unlock the potential of genomic data to provide a reliable estimate of past human census size, hence informing other historical sciences such as the study of cultural evolution, history and archaeology.

METHODS

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Whole Genome Sequencing data

We used whole genome sequences of the Estonian Biobank participants reported in Kals et al.[10]. We applied exactly the same criteria for filtering individuals (sequencing quality control filters, match between WGS and chip genotype, total number of SNVs, self-reported Estonian ethnicity etc., see Kals et al.[10] for details) except for relatedness and singleton count (see below). For manipulations with vcf files bcftools-1.8[36] was used unless specified otherwise. Additionally to sample filtering applied by Kals et al.[10], we removed seven samples with missing call over 3% as well as related individuals. To do so we used PLINK-1.9[37] and KING-2.1.6[38] to estimate relatedness coefficient and removed one individual from each pair with this value equal to 0.0442 or higher, corresponding to third degree relatives. This resulted in a dataset consisting of 2,305 individuals that was used for all downstream analyses.

For analyses that require phased and/or imputed data (CHROMOPAINTER, SDS) phasing and imputation was done using Eagle v2.3[39] on the dataset consisting of 2,420 samples to

benefit from the presence of related individuals and subsequently relevant samples were extracted.

All Estonian Biobank participants have signed a broad informed consent which allows research in the fields of genetic epidemiology, disease risk factors and population history. All work at Estonian Biobank is conducted according to the Estonian Human Gene Research Act. The original study generating the WGS data[10] was approved by the Research Ethics Committee of the University of Tartu (application number 234/T-12).

The 'Rural above 50 years old' (R50+) panel

As information on parents' and grandparents' birthplace is mostly unavailable for the samples used here, we subsetted the 2,305 dataset for individuals born in rural areas and sampled at the age of 50 or older as we expect this cohort to be the least affected by recent migration and long-distance marriages, hence expecting it to preserve the original genetic structure. This resulted in a dataset of 474 individuals which we further pruned for PCA outliers (see below) and samples with more than 10,000 singletons (SI1:1.1-SI1:1.3). Per-sample number of singletons was estimated using vcftools-0.1.14[40] on the entire (2,305 samples) non-imputed dataset. We ended up with a panel of 468 individuals, which we call "R50+".

Non-Estonian samples

To place the Estonian population genetic variation in Eurasian context we compiled two datasets containing the R50+ Estonian samples each and samples from various populations predominantly representing West Eurasia. The first dataset used for PCA contained 59 samples from 17 populations sequenced on the Complete Genomics platform, 207 samples from 8 populations sequenced using Illumina technology and 255 samples from 14 populations genotyped on Illumina arrays (Table SI2-1.2-I). Whole genome sequences were pruned to keep positions matching those overlapping between genotyping arrays resulting in approximately 450K SNPs.

The second dataset used for CHROMOPAINTER/fineSTRUCTURE/GLOBETROTTER except for R50+ Estonians included 425 samples from 27 populations genotyped on Illumina arrays and 175 samples from seven 1000 Genome Project populations (CHB, FIN, GBR, GIH, IBS, TSI, YRI) (Table SI2-3.1-I). Whole genome sequences were pruned to keep positions matching those overlapping between genotyping arrays resulting in approximately 500K SNPs.

Principal component analysis

We ran principal component analysis (PCA) for the entire Estonian dataset in two settings: **a)** with only the 2,305 Estonians and **b)** combining the 2,305 Estonians with 521 non-Estonian samples from 18 European populations (Table SI2-1.2-I). In both cases *smartPCA* from EIGENSOFT-7.2.0[41] was used. In setting **a** we directly ran PCA on the dataset filtering for MAF below 0.01, no-call above 0.03 and positions in LD ($r^2 > 0.4$ within sliding windows of 200 positions). Results obtained in setting **a** were used to identify Estonian samples with extreme position in the PCA plot to be removed from the R50+ panel and from the dataset used for SDS (S1:1.2). In setting **b** we used 255,536 bi-allelic SNPs that overlap between the different datasets and passed LD-pruning ($r^2 > 0.6$ within sliding windows of 1,000 positions), MAF (<0.05) and no-call (>0.05) filters. We first calculated the principal components (PCs) based on all non-Estonian samples and then projected the Estonian individuals onto the first two PCs.

CHROMOPAINTER/fineSTRUCTURE/GLOBETROTTER

To study genetic similarities between Estonians and other European populations we used the CHROMOPAINTER/fineSTRUCTURE (CP/FS) pipeline[17]. It involves a chromosome "painting" procedure which represents each chromosome of an individual (the recipient) as a mixture of chunks received (copied) from every other individual in the dataset (donor). The number of chunks copied by a recipient from each of the donors makes a "copying vector" which are used in the FS algorithm to group individuals into populations.

Initial chromosome painting parameters were estimated using 30% of the phased dataset of 1068 Estonian and non-Estonian samples (Table SI2-3.1-I). FS was run for 15 million MCMC iterations in two parallel runs to assess convergence. The tree-building step was performed using the approach from Leslie et al.[1] and the run with the highest observed posterior likelihood was used to cluster the samples into genetic groups. Inferred FS groups were further manually inspected and clustered into the higher-order FS populations (S1:3.2). These FS groups were used as surrogate populations to infer admixture with GLOBETROTTER and estimate ancestry profile with NNLS.

Next, GLOBETROTTER (GT)[17] was used to detect signals and dates of admixture for the Estonian groups defined using the approach described above. Unlike many other methods, GT allows the structure of unsampled source populations which were involved in the admixture event(s) to be assessed by modelling them as a mixture of sampled surrogate populations. GT inference was performed using a "regional" approach[17,42]. Estonian clusters were only allowed to copy from external surrogates, but not from other Estonians. CHROMOPAINTER parameters were estimated for each Estonian target group individually and the average over all target populations was used to prepare input copying vectors for GT. Two separate runs, with and without standardization by "NULL" individual, were performed and consistency between runs was checked. To assess whether unbalanced surrogate population sample size could have biased our GT inference, we performed five additional GT runs by down sample both target and surrogate populations to 20 individuals.

Finally, given complex admixture signal in Estonia, we implemented non-negative least-squares (NNLS) method[1]. This allowed us to assign relative ancestral proportions to each individual in the R50+ panel using the non-Estonian surrogate groups identified by FS as sources. NNLS values for CP/FS Estonian groups were extracted from GT output while for individual samples these were calculated with an in-house R script. Obtained results were then summarized across Estonian parishes as well as across IBD/FS clusters.

Detecting segments identical-by-descent (IBD segments)

To detect IBD segments in the Estonian dataset we applied *IBDseq* version r1206 (*10*) with default settings (errormax=0.001, errorprop=0.25, r2window=500, r2max=0.15, minalleles=2, lod=3.0) to the non-phased non-imputed dataset consisting of 2,305 Estonians. Choosing *IBDseq* over *refined IBD*[43] here is justified by working with samples coming from a relatively homogeneous population, which makes *IBDseq* frequency model applicable, while *IBDseq* has the advantage of not requiring phasing as well as having sequencing errors and rare alleles being explicitly accounted for. As *IBDseq* software reports only physical coordinates of a segment's start and end we interpolated segments' genetic length in cM using GRCh37 recombination map (ftp://ftp.ncbi.nlm.nih.gov/hapmap/recombination/2011-01_phaseII_B37/) using R[44]. When working with the R50+ panel corresponding IBD segments were retrieved from the general output obtained on the 2,305 dataset. Homozygosity-by-descent segments were also inferred with *IBDseq*.

IBD segments between Estonians and non-Estonian individuals were detected by applying *refined IBD* version 12Jul18.a0b[43] with default parameters except for length (window=40.0, length=1.0, trim=0.15, lod=3.0) to the same dataset that was used for CP/FS/GT, as in this case the dataset is highly structured. This was followed by applying the *merge-ibd* utility version 12Jul18.a0b to merge together segments separated by at most 1 cM long gaps and no more than 2 positions with genotypes discordant with IBD.

Both for *IBDseq* and *refined IBD/ibd-merge* results segments shorter than 2 cM were discarded, as longer segments are detected with higher reliability.

MAPS

In order to evaluate the extent of gene flow across the whole country together with local population densities, we estimated migration surfaces using MAPS[15], which harnesses a similarity matrix summarizing the total number of IBD segments shared in a given population. In doing so, we used the IBD segments shared among pairs of individuals inferred with *IBDseq* as described in the previous section. Subsequently we have classified the shared genetic fragments as "short" (between 2 and 6 cM) and "long" (more than 6 cM),

and performed two independent MAPS runs for each length class to assess convergence. Estonian territory was modeled as having a total of 200 demes. Each run had 5 million iterations thinned every 10,000 and preceded by a burn-in of 2 million discarded cycles. The obtained migration surfaces were subsequently plotted using the plotmaps R package[15]. We repeated the whole procedure after removing samples belonging to clusters from Figure 2 with mean sharing above 60 cM to assess their effect on MAPS results.

IBD-based fineSTRUCTURE (IBD/FS)

In order to exploit patterns of genetic similarity between samples that arose very recently and get insight into fine genetic structure of the Estonian population, we used total genetic length of IBD segments longer than 2 cM as a measure of genetic similarity between pairs of individuals. We refer to this measure as “IBD-sharing”. Next, to obtain natural genetic grouping of the samples we used a matrix of IBD-sharing as input for fineSTRUCTURE v2.0.7[14]. Although our approach is different from the original CHROMOPAINTER/fineSTRUCTURE method[14], it is very similar in its idea to the approach used in [3], and, put loosely, treats each cM shared between a pair of individuals as a CHROMOPAINTER chunk copied by the recipient from the donor. The fineSTRUCTURE algorithm already has an inbuilt method of compensating for the fact that the units used to measure similarity/relatedness between samples (either chunks in the classical approach or cM in our approach) don’t represent fully independent pieces of information by transforming the raw value into an effective one by applying a *c*-factor. The *c*-factor was calculated using the *fs combine* command with the -C option applied to matrices of IBD-sharing for each individual chromosome. For more details see Supplementary Information SI1:2.1 and SI1:2.2. When running fineSTRUCTURE for both R50+ and the entire dataset the first 2,000,000 MCMC iterations were removed as burn-in and subsequently MCMC was run for additional 2,000,000 MCMC iterations sampling every 10,000th run. When building the tree we used the approach described in Leslie et al., 2015[1] and corresponding to the “1” value of the -T option in fineSTRUCTURE v2, which, put informally, maximizes the concordance

between samples' final cluster assignment and its' assignment in individual MCMC runs. To validate this approach we applied it to the simulated data used in Lawson et al., 2012[14], and calculated the same measure of correlation between real and inferred cluster assignment of the samples for different number of chromosomal regions used to detect IBD segments (SI1:2.2).

We applied this approach to the R50+ dataset (468 samples) and the entire dataset (2,305 samples). In both cases fineSTRUCTURE was run twice to assess convergence (SI1:2.3, Tables SI2-2.3-I and SI2-2.3-II). In the case of the R50+ dataset to reduce the number of clusters revealed by the fineSTRUCTURE algorithm we have hierarchically joined together clusters with short terminal branches by cutting the tree at such a level so as to avoid having clusters consisting of less than 5 samples. In the case of the entire dataset clusters referred to throughout the text were obtained by cutting the tree at a level chosen after visual inspection (SI1:2.3).

Fst calculations

Fst between Estonian clusters was calculated using smartpca from the EIGENSOFT package v7.2.0[41] after LD-pruning ($r^2 > 0.4$, windows of 1,000 SNPs) and removing sites with MAF < 0.05 and missing rate > 0.1.

Per-site Weir and Cockerham[32] Fst estimator between SE and nonSE subsets was calculated using vcfttools[40] after filtering sites for LD, MAF and missing rate the same way as described above.

Geographic data visualization

Geographic coordinates of the corresponding birth town/parish were assigned to each sample with birth place information available (2,168 out of 2,305 samples). For MAPS these coordinates were used directly. When plotting IBD/FS and NNLS results for the R50+ panel, coordinates of the samples were changed manually to avoid over-plotting. When plotting samples from the entire dataset jittering was used for the same purpose. Shp objects used to plot maps of Estonia with parish and county borders were retrieved from the Estonian

Land Board website (Administrative and settlement units, 2018.11.01, https://geoportaal.maaamet.ee/eng/Spatial-Data/Administrative-and-Settlement-Division-p312.html). Geographic data were visualized in R[44] with the aid of the following packages: sp[45,46], sf[47], rgdal[48], rgeos[49] and ggplot2[50].

IBDNe

In order to reconstruct recent Ne dynamics we used *IBDNe* version 07May18.6a4[19] with default settings (npairs=0, nits=1000, nboots=80, trimcm=0.2). *IBDNe* was applied to sets of no less than 100 individuals. In all cases IBD segments used as input for *IBDNe* were detected with *IBDseq*[13]. Recombination map in PLINK format used to convert physical distances to the genetic ones was taken from http://bochet.gcc.biostat.washington.edu/beagle/genetic_maps/plink.GRCh37.map.zip.

To get independent evidence of regional differences in Ne dynamics we applied *IBDNe* to samples from the People of the British Isles[1] dataset grouped by the region of origin of individuals' grandparents. The following regions were used: Scotland, Wales and North-East, North-West, South-East and South-West England. For the list of counties comprising these regions see Table S2:4.3-I.

Genetic simulations

To simulate genetic data under various demographic scenarios to test the behavior of IBDNe we used mspms which is an ms-compatible version of msprime[51]. Commands used for simulation are provided in the Supplementary Information S1:4.1.

Estimating actual census size based on Ne

Several lines of evidence, based both on theoretical reasoning[52] and empirical comparisons[19] suggest that in industrial human societies census size (Nc) is roughly 3 fold the Ne assuming a panmictic and isolated population. However, application of this coefficient is limited to populations with specific reproductive characteristics, for example 1:1 male to female ratio and Poisson distribution of number of offspring among individuals capable of

reproducing. We therefore adapted a more general approach from[22] which incorporates the inbreeding coefficient (Fis), relative fraction of males (m) and excess in variance of reproductive success compared to the Poisson distribution (DV):

$$N_{b(t)} = \frac{(1+Fis)}{4} \times \left(\frac{1}{(1-m)\times m} + DV\right) \times N_{e(t)} \quad [1]$$

Formula [1] yields the N_b, the number of breeding individuals (individuals capable of reproducing) at time t under the assumption of absence of gene flow and population structure, non-overlapping generations and equal variance of number of offspring between sexes. It is dependent on parameters such as m, Fis and DV that cannot be reliably estimated for each time bin. We therefore explored a range of plausible scenarios described by different values of m, Fis and DV based on the following assumptions: i) Fis calculated on chromosome 1 for contemporary human populations from the 1000 genomes dataset[53] as well as for Estonians ranges from −0.016 to 0.004 (S1:4.4), leading to the conclusion that for most human populations the term (1+Fis) can be safely approximated to 1; ii) m, the relative fraction of reproducing males, must be comprised between 0.1 and 0.9, considering further polarizations of this parameter as implausible for our species; iii) DV, the difference between the expected and the observed variance in number of offspring per adult can be estimated to range between −1 and 3. The latter estimate was obtained by taking Poisson distributions constrained between 0 and 10 (considering 10 as the maximum number of surviving children per adult) with an average between 1 and 5, and by empirically inflating the most extreme bins (0 or 10 children per adult) 5-fold. Such an exercise yields DVs ranging between −0.2 and 2.5, which we conservatively rounded to −1 and 3, respectively. This range is also consistent with data from contemporary Estonians, available from the Estonian Biobank, and showing a DV of −0.76 based on 7,863 females born between 1900 and 1955 and in the age of menopause at the time of enrolment. When plugged together into formula [1], these estimates yield a minimum of 0.75 (with m=0.5 and DV=−1) and a maximum of 3.53 (with m=0.1 or 0.9 and DV=3). To provide a single point estimate of Nc we rewrite formula [1] as

$$N_{b(t)} = 1.63 \times N_{e(t)} \quad [2]$$

using a geometric mean between 0.75 and 3.5 and thus making our estimate slightly more than 2-fold away from the provided range boundaries. Note, that although there are indications that in some human populations DV can be higher than 3[54], such cases can be considered to be at the very extreme of human reproductive behaviour spectrum as even hypothetical “super-male” populations would have a sex-average DV of 1.8 given m equals to 0.5[55]. Hence we suggest our approach to be applicable to many human populations provided that immigration and population structure can be properly accounted for. In addition, the range of DV can be changed to study populations with extreme inequality in reproductive success.

The value estimated using [2] corresponds to the number of individuals in reproductive age. It can be converted into total census size (Nc) of a human population at a given time point by dividing it by the estimated fraction of breeding individuals, which we here assume to be roughly 0.33. This is supported by actual data on the Estonian population from the “Statistics Estonia” database (http://andmebaas.stat.ee/Index.aspx?lang=en#) showing that the fraction of people between 20 and 40 years old was between 0.33 to 0.38 during the period between 1970 and 2018. Incorporating this idea into [2] results in equation [3].

$$N_{c(t)} = 4.89 \times N_{e(t)} \quad [3]$$

which we used to obtain the curve in Figure 6B. Sources of historical estimates of Estonian population size used in that figure are provided in S4.2-II.

When using Ne as a proxy for actual population size one should keep in mind the potential effect of gene flow between populations. For example under a stepping stone model with constant population size and migration Ne estimated using samples from one deme is expected to increase when going back in time as more and more ancestors of sampled individuals would represent other demes[56]. In other words, coalescent-based Ne estimates

reflect the number of ancestors of a sampled population, which may have lived in any location in space, rather than strictly the number of individuals in a given area at a given time point.

Singleton density score (SDS) selection scan

As SDS[20] does not handle missing data, imputed genomes of 2,301 unrelated individuals (four PCA outliers removed) were used. SDS[20] analysis was applied to three datasets separately, namely, the entire dataset and it's two subsets, Estonia SE and Estonia nonSE. The latter two were defined based on the IBD/FS results (Figure 5): SE (individuals with South-East Estonian ancestry belonging to clusters eSE_1-eSE_5) and nonSE (individuals coming from the other parts of the country and belonging to other clusters). We did not apply SDS to finer subclusters of the dataset due to sample size issues. The number of individual genomes used in the analysis has a direct impact on the number of SNPs analyzed, the power to detect selection at any given SNP and the length of the terminal branches of the coalescent tree and hence the timing of the selection events that can be inferred[20]. Since in this study we are focusing on recent selection signals and differences between subpopulations that show only limited differentiation, we aimed at using samples of at least 1000 individuals. As the dataset needs to be homogeneous in terms of number of singletons per individual, this value was calculated with vcftools 0.1.14[40] independently for each of the datasets and individuals with extreme values (below 5th or above 95th quantiles) were removed. Final datasets included 2,076, 927 and 1,132 samples for the entire dataset, SE and nonSE subsets respectively. Predicted functional effect of the test SNPs was assessed using Combined Annotation-Dependent Depletion tool (CADD)[29]. In addition, two alternative enrichment tests were performed to see whether candidate SNPs are enriched in a certain category of genes[30,31] or in certain GWAS catalogue categories (http://www.ebi.ac.uk/gwas/home;[57]). Candidate SNPs, as well as SNPs in linkage disequilibrium with those (R2 > 0.5), were checked for known e-QTL effects using the

eQTLGen Consortium[58] (http://www.eqtlgen.org/) database. Details of SNP annotation and enrichment analyses are specified in S1:5.2.

DATA AVAILABILITY

The sequencing data are available on demand. The procedure of applying for the access to the data can be found under the following link: https://www.geenivaramu.ee/en/biobank.ee/data-access.

REFERENCES

1 Leslie S, Winney B, Hellenthal G *et al.* The fine-scale genetic structure of the British population. *Nature* 2015; **519**: 309–314.

2 Martin AR, Karczewski KJ, Kerminen S *et al.* Haplotype Sharing Provides Insights into Fine-Scale Population History and Disease in Finland. *Am J Hum Genet* 2018; **102**: 760–775.

3 Bycroft C, Fernandez-Rozadilla C, Ruiz-Ponte C *et al.* Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula. *Nat Commun* 2019; **10**: 551.

4 Raveane A, Aneli S, Montinaro F *et al.* Population structure of modern-day Italians reveals patterns of ancient and archaic ancestries in Southern Europe. *Sci Adv* 2019; **5**: eaaw3492.

5 Saint Pierre A, Giemza J, Alves I *et al.* The genetic history of France. *Eur J Hum Genet* 2020; : 1–13.

6 Berg JJ, Harpak A, Sinnott-Armstrong N *et al.* Reduced signal for polygenic adaptation of height in UK Biobank. *eLife* 2019; **8**: e39725.

5227 Sohail M, Vakhrusheva OA, Sul JH *et al.* Negative selection in humans and fruit flies
523 involves synergistic epistasis. *Science* 2017; **356**: 539–542.

5248 Haworth S, Mitchell R, Corbin L *et al.* Apparent latent structure within the UK Biobank
525 sample has implications for epidemiological analysis. *Nat Commun* 2019; **10**: 333.

5269 Kerminen S, Martin AR, Koskela J *et al.* Geographic Variation and Bias in the Polygenic
527 Scores of Complex Diseases and Traits in Finland. *Am J Hum Genet* 2019; **104**: 1169–
528 1181.

52910 Kals M, Nikopensius T, Läll K *et al.* Advantages of genotype imputation with ethnically
530 matched reference panel for rare variant association analyses. *bioRxiv* 2019; : 579201.

53111 Nelis M, Esko T, Mägi R *et al.* Genetic Structure of Europeans: A View from the North–
532 East. *PLOS ONE* 2009; **4**: e5472.

53312 Haller T, Leitsalu L, Fischer K *et al.* MixFit: Methodology for Computing Ancestry-Related
534 Genetic Scores at the Individual Level and Its Application to the Estonian and Finnish
535 Population Studies. *PLoS ONE* 2017; **12**. doi:10.1371/journal.pone.0170325.

53613 Browning BL, Browning SR. Detecting Identity by Descent and Estimating Genotype
537 Error Rates in Sequence Data. *Am J Hum Genet* 2013; **93**: 840–851.

53814 Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of Population Structure using
539 Dense Haplotype Data. *PLoS Genet* 2012; **8**: e1002453.

54015 Al-Asadi H, Petkova D, Stephens M, Novembre J. Estimating recent migration and
541 population-size surfaces. *PLOS Genet* 2019; **15**: e1007908.

54216 Kallio P. The Diversification of Proto-Finnic. In: Ahola J, Frog, Tolley C (eds). *Fibula,*
543 *Fabula, Fact: The Viking Age in Finland*. BoD - Books on Demand, 2018.

54417 Hellenthal G, Busby GBJ, Band G *et al.* A genetic atlas of human admixture history.
545 *Science* 2014; **343**: 747–751.

54618 Loit A. Invandringen från Finland till Baltikum under 1600-talet. *Hist Tidskr För Finl* 1982;
547 : 194–195.

54819 Browning SR, Browning BL. Accurate Non-parametric Estimation of Recent Effective
549 Population Size from Segments of Identity by Descent. *Am J Hum Genet* 2015; **97**: 404–
550 418.

55120 Field Y, Boyle EA, Telis N *et al.* Detection of human adaptation during the past 2000
552 years. *Science* 2016; **354**: 760–764.

55321 Laporte V, Charlesworth B. Effective Population Size and Population Subdivision in
554 Demographically Structured Populations. *Genetics* 2002; **162**: 501–519.

55522 Charlesworth B. Fundamental concepts in genetics: effective population size and
556 patterns of molecular evolution and variation. *Nat Rev Genet* 2009; **10**: 195–205.

55723 Li H, Durbin R. Inference of human population history from individual whole-genome
558 sequences. *Nature* 2011; **475**: 493–496.

55924 Okada Y, Momozawa Y, Sakaue S *et al.* Deep whole-genome sequencing reveals recent
560 selection signatures linked to evolution and disease risk of Japanese. *Nat Commun*
561 2018; **9**. doi:10.1038/s41467-018-03274-0.

56225 Theisen DJ, Davidson JT, Briseño CG *et al.* WDFY4 is required for cross-presentation in
563 response to viral and tumor antigens. *Science* 2018; **362**: 694–699.

56426 Yuan Q, Li Y, Li J *et al.* WDFY4 Is Involved in Symptoms of Systemic Lupus
565 Erythematosus by Modulating B Cell Fate via Noncanonical Autophagy. *J Immunol*
566 *Baltim Md 1950* 2018; **201**: 2570–2578.

Zhang Y, Bo L, Zhang H, Zhuang C, Liu R. E26 Transformation-Specific-1 (ETS1) and WDFY Family Member 4 (WDFY4) Polymorphisms in Chinese Patients with Rheumatoid Arthritis. *Int J Mol Sci* 2014; **15**: 2712–2721.

McIntosh LA, Marion MC, Sudman M *et al.* Genome-Wide Association Meta-Analysis Reveals Novel Juvenile Idiopathic Arthritis Susceptibility Loci. *Arthritis Rheumatol Hoboken NJ* 2017; **69**: 2222–2232.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014; **46**: 310–315.

Chen EY, Tan CM, Kou Y *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 2013; **14**: 128.

Kuleshov MV, Jones MR, Rouillard AD *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 2016; **44**: W90-97.

Weir B, Clark Cockerham C. Weir BS, Cockerham CC.. Estimating F-Statistics for the Analysis of Population-Structure. Evolution 38: 1358-1370. *Evolution* 1984; **38**: 1358–1370.

Minassian BA, Lee JR, Herbrick JA *et al.* Mutations in a gene encoding a novel protein tyrosine phosphatase cause progressive myoclonus epilepsy. *Nat Genet* 1998; **20**: 171–174.

Serratosa JM, Gómez-Garre P, Gallardo ME *et al.* A novel protein tyrosine phosphatase gene is mutated in progressive myoclonus epilepsy of the Lafora type (EPM2). *Hum Mol Genet* 1999; **8**: 345–352.

Nitschke F, Ahonen SJ, Nitschke S, Mitra S, Minassian BA. Lafora disease - from pathogenesis to treatment strategies. *Nat Rev Neurol* 2018; **14**: 606–617.

59136 Li H, Handsaker B, Wysoker A *et al.* The Sequence Alignment/Map format and
592 SAMtools. *Bioinforma Oxf Engl* 2009; **25**: 2078–2079.

59337 Purcell S, Neale B, Todd-Brown K *et al.* PLINK: a tool set for whole-genome association
594 and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.

59538 Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. Robust
596 relationship inference in genome-wide association studies. *Bioinformatics* 2010; **26**:
597 2867–2873.

59839 Loh P-R, Palamara PF, Price AL. Fast and accurate long-range phasing in a UK Biobank
599 cohort. *Nat Genet* 2016; **48**: 811–816.

60040 Danecek P, Auton A, Abecasis G *et al.* The variant call format and VCFtools. *Bioinforma*
601 *Oxf Engl* 2011; **27**: 2156–2158.

60241 Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet*
603 2006; **2**: e190.

60442 Hudjashov G, Karafet TM, Lawson DJ *et al.* Complex Patterns of Admixture across the
605 Indonesian Archipelago. *Mol Biol Evol* 2017; **34**: 2439–2452.

60643 Browning BL, Browning SR. Improving the Accuracy and Efficiency of Identity-by-
607 Descent Detection in Population Data. *Genetics* 2013; **194**: 459–471.

60844 R Core Team. *R: A language and environment for statistical computing*. R Foundation for
609 Statistical Computing: Vienna, Austria, 2018https://www.R-project.org/.

61045 Pebesma E, Bivand R. Classes and methods for spatial data in R. *R News* 2005; **5**: 9–
611 13.

61246 Bivand RS, Pebesma E, Gómez-Rubio V. *Applied Spatial Data Analysis with R*. 2nd ed.
613 Springer-Verlag: New York, 2013https://www.springer.com/gp/book/9781461476177
614 (accessed 18 Jun2019).

61547 Pebesma E. Simple Features for R: Standardized Support for Spatial Vector Data. *R J*
616 2018.https://journal.r-project.org/archive/2018/RJ-2018-009/.

61748 Bivand R, Keitt T, Rowlingson B *et al. rgdal: Bindings for the 'Geospatial' Data*
618 *Abstraction Library*. 2019https://CRAN.R-project.org/package=rgdal (accessed 18
619 Jun2019).

62049 Bivand R, Rundel C, Pebesma E *et al. rgeos: Interface to Geometry Engine - Open*
621 *Source ('GEOS')*. 2019https://CRAN.R-project.org/package=rgeos (accessed 18
622 Jun2019).

62350 Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag: New York,
624 2009https://www.springer.com/gp/book/9780387981413 (accessed 18 Jun2019).

62551 Kelleher J, Etheridge AM, McVean G. Efficient Coalescent Simulation and Genealogical
626 Analysis for Large Sample Sizes. *PLOS Comput Biol* 2016; **12**: e1004842.

62752 Felsenstein J. Inbreeding and Variance Effective Numbers in Populations with
628 Overlapping Generations. *Genetics* 1971; **68**: 581–597.

62953 The 1000 Genomes Project Consortium. A global reference for human genetic variation.
630 *Nature* 2015; **526**: 68–74.

63154 Austerlitz F, Heyer E. Social transmission of reproductive behavior increases frequency
632 of inherited disorders in a young-expanding population. *Proc Natl Acad Sci U S A* 1998;
633 **95**: 15140–15144.

Heyer E, Chaix R, Pavard S, Austerlitz F. Sex-specific demographic behaviours that shape human genomic variation. *Mol Ecol* 2012; **21**: 597–612.

Browning SR, Browning BL, Daviglus ML *et al.* Ancestry-specific recent effective population size in the Americas. *PLoS Genet* 2018; **14**: e1007385.

MacArthur J, Bowler E, Cerezo M *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* 2017; **45**: D896–D901.

Võsa U, Claringbould A, Westra H-J *et al.* Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv* 2018; : 447367.

Author Contribution

VP, LP, AM and MM designed the study. LS, TE, RM, LP and AM conducted sample management and provided access to data. MS provided historical data. VP, FM, AK, GH, FJ, RF, DM, MK, AA, DJL and LP analyzed the data. VP, FM, AK, GH, FJ, LS, RF, DM, MS, UV, MK, CT, MMo, LM, AA, DJL, TE, RM, LP, AM, MM contributed to the interpretation of results. VP, FM, AK, GH, FJ, RF, DM, MS, CT, DJL, LP, MM wrote the manuscript.

Acknowledgements

This work was supported by the Estonian Research Council grants PRG243 (GH, RF, LS, LP, MM), PUT1339 (AK), PUTJD817 (MK), MOBTP108 (UV), IUT20-60 (AM, RM), IUT24-6 (AM, RM), PUT1660 (TE) and PRG184 (LM); by the European Union through the European Regional Development Fund Projects No. 2014-2020.4.01.16-0030 (FM, MM, VP), No. 2014-2020.4.01.15-0012 (AM, RM, TE, LM, GH, LS, MM), No. 2014-2020.4.01.16-0024 (DM, LP, VP), No. 2014-2020.4.01.16-0125 (RF), No. 2014-2020.4.01.16-0271 (RF) and 2014-2020.4.01.16-0125 (AM, RM, TE, LM); by NIH GIANT (AM, TE); by ERA-CVD grant

657Detectin-HF (AM), by NIH-BMI Grant 2R01DK075787-06A1 (TE), by the Wellcome Trust No. 658WT104125MA (DJL) as well as by the European Union through Horizon 2020 grant no. 659810645 (MM) and PRESICE4Q (LM).

660Data analysis was performed on the High-Performance Computing Center of University of 661Tartu.

662The authors would like to thank the Genome Aggregation Database (gnomAD) and the 663groups that provided exome and genome variant data to this resource. A full list of 664contributing groups can be found at https://gnomad.broadinstitute.org/about.

665The authors would like to thank Bayazit Yunusbayev for fruitful discussion and valuable 666advice.

667**Competing interests**

668Authors declare no competing interests.

669**Figures**

670Figure 1. Principal components analysis of 2,305 Estonian samples in the context of West 671Eurasian populations. Estonian samples were projected onto PC space defined by European 672samples (Methods, SI1:1.1, SI1:1.2, Table SI2:1.1-I, Table SI2:1.2-I). Outlined labelled dots 673correspond to medians of European populations or Estonian counties while non-outlined 674dots represent individual samples. Estonian samples are shown in colour corresponding to 675the geographic region of origin: NW (North-West) included Harjumaa (Ha), Läänemaa (Lä), 676Raplamaa (Ra), Järvamaa (Jä), Hiiumaa (Hi) and Saaremaa (Sa) counties; NE (North-East) 677includes Lääne-Virumaa (LV), Ida-Virumaa (IV) and Jõgevamaa (Jõ) counties; SW (South-678West) includes Pärnumaa (Pä) and Viljandimaa (Vi); SE (South-East) includes Valgamaa 679(Va), Tartumaa (Ta), Põlvamaa (Põ) and Võrumaa (Võ). NA are individuals with no birth

680place information available. Individual non-Estonian samples are shown in grey. Medians of
681non-Estonian populations are coloured according to geographic regions as shown in the
682legend. Russians (N) and Russians (C/S) refers to Russians North and Central/South (Table
683SI2:1.2-I). Inset in bottom left corner shows a map of Estonian counties. This map was
684created in R (https://www.R-project.org/)44 using an shp object of the Administrative and
685settlement units provided by the Estonian Land Board, 2018.11.01
686(https://geoportaal.maaamet.ee/eng/Spatial-Data/Administrative-and-Settlement-Division-
687p312.html). See Methods for more details.

688Figure 2. Genetic clustering of R50+ samples based on pairwise sharing of IBD segments. a,
689Hierarchical relationships (tree) and the average total length of IBD segments shared
690between cluster members (heatmap) as inferred by fineSTRUCTURE. The length of the tree
691branches does not reflect any relationship between the clusters. Clusters are named to
692reflect their geographic distribution (E - “East”, NW - “North-West”, NE - “North-East”, SW -
693“South-West”, SE - “South-East”). Numbers in grey next to cluster names refer to the sample
694size of each cluster. b, Geographic distribution of inferred genetic clusters. Each symbol on
695the Estonian map corresponds to one individual from the R50+ subset. See SI1:2.3 for
696details. This map was created in R (https://www.R-project.org/)44 using an shp object of the
697Administrative and settlement units provided by the Estonian Land Board, 2018.11.01
698(https://geoportaal.maaamet.ee/eng/Spatial-Data/Administrative-and-Settlement-Division-
699p312.html). See Methods for more details.

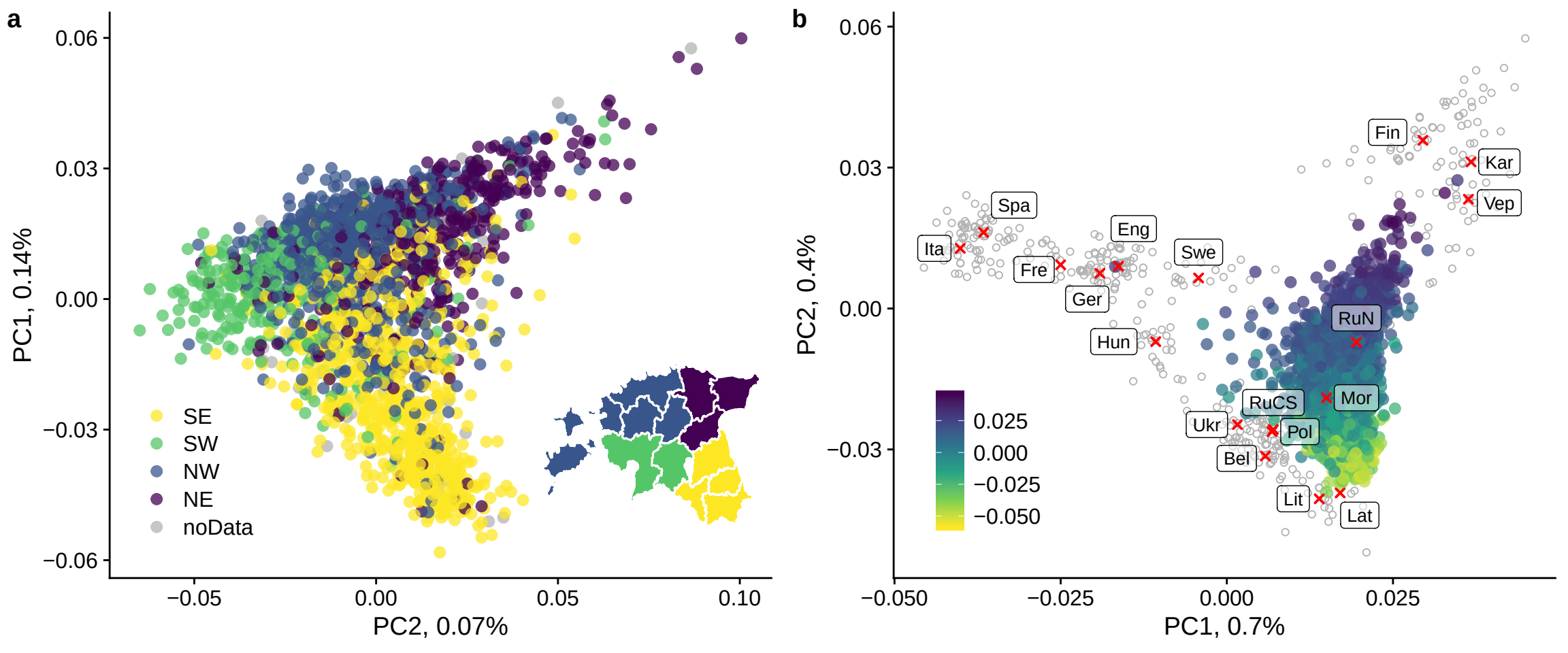
700Figure 3. Homozygosity-by-descent in the R50+ dataset. Boxplots show distribution of per-
701genome total length of Homozygosity-By-Descent (HBD) tracts within clusters shown in
702Figure 2. HBD tracts were detected using IBDseq13. The boxes show 25th, 50th and 75th
703quantiles, while the whiskers show values within 1.5 times the inter-quantile range (IQR) of
70425th and 75th quantiles. Individual dots show outliers (values out of the range shown by the
705whiskers).

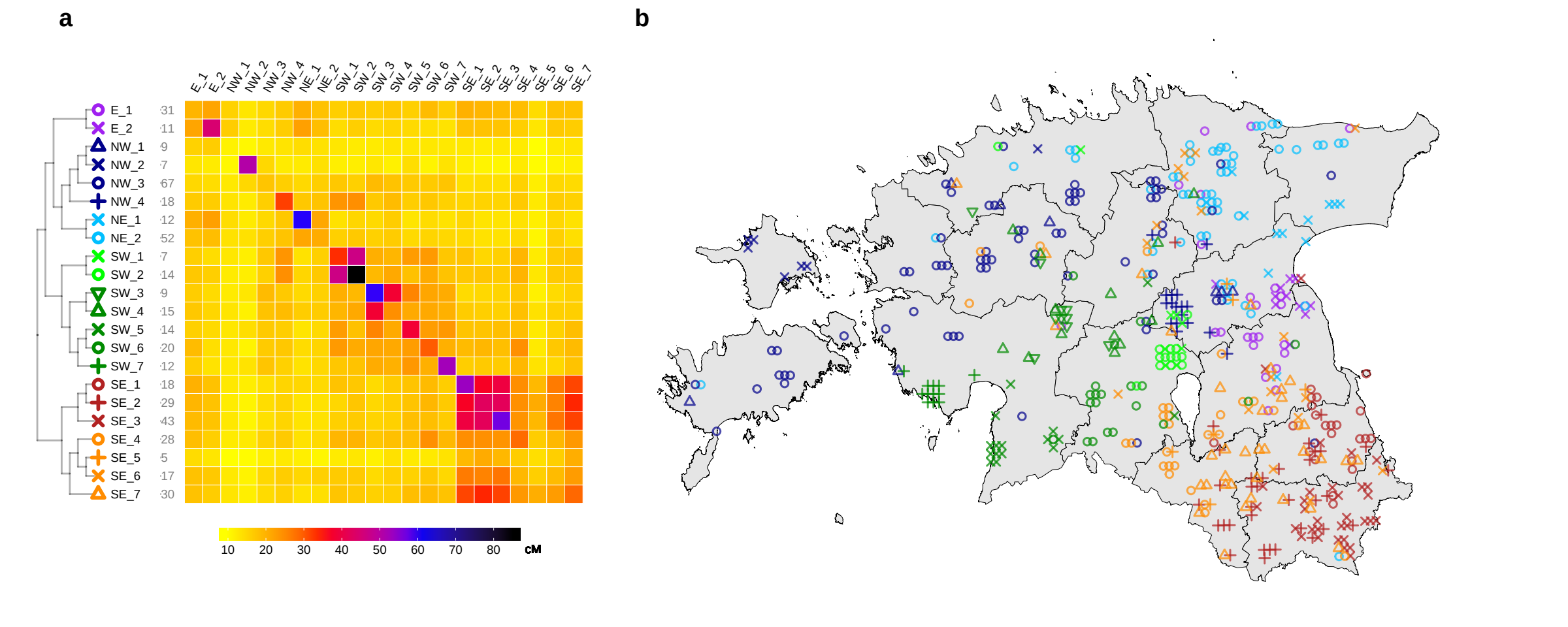
706Figure 4. Relative proportions of “Baltic”, “Slavic”, Finnish and Swedish ancestry in the R50+
707subset. Modelled relative ancestral proportions of «Balts» (Latvians and Lithuanians) (a),
708«Slavs» (Belarusians, Poles, Russians, Ukrainians) (b), Finns (c), and Swedes (d) attributed
709by applying non-negative least squares approach (NNLS) to
710CHROMOPAINTER/fineSTRUCTURE (CP/FS) results are shown. For details on source
711group composition, as well as for results for other groups see SI1:3.1. The colour of each
712parish reflects mean values of samples coming from this parish. Parishes with no samples in
713the R50+ dataset are filled with grey. The scale is the same for all four panels. See SI1:3.1
714for more details. These maps were created in R (https://www.R-project.org/)44 using an shp
715object of the Administrative and settlement units provided by the Estonian Land Board,
7162018.11.01 (https://geoportaal.maaamet.ee/eng/Spatial-Data/Administrative-and-Settlement-
717Division-p312.html). See Methods for more details.

718Figure 5. Genetic clustering of the entire Estonian dataset (2,305 samples). Samples were
719clustered using the fineSTRUCTURE clustering algorithm based on pairwise total genetic
720length of IBD segments as described in Methods. Obtained clusters were pulled together
721based on their position on the tree resulting in 12 higher order clusters shown here (SI1:2.3).
722A: Hierarchical relationships (tree) and average total length of IBD segments shared
723between clusters (heatmap). The length of the tree branches does not reflect any
724relationship between the clusters. Numbers in grey next to cluster names show the number
725of samples in each cluster. B: Geography of inferred clusters. Each dot within the contour of
726Estonia corresponds to one individual, while waffle plots show samples for 15 major
727Estonian towns with each dot corresponding to 5 individuals. This map was created in R
728(https://www.R-project.org/)44 using an shp object of the Administrative and settlement units
729provided by the Estonian Land Board, 2018.11.01
730(https://geoportaal.maaamet.ee/eng/Spatial-Data/Administrative-and-Settlement-Division-
731p312.html). See Methods for more details.

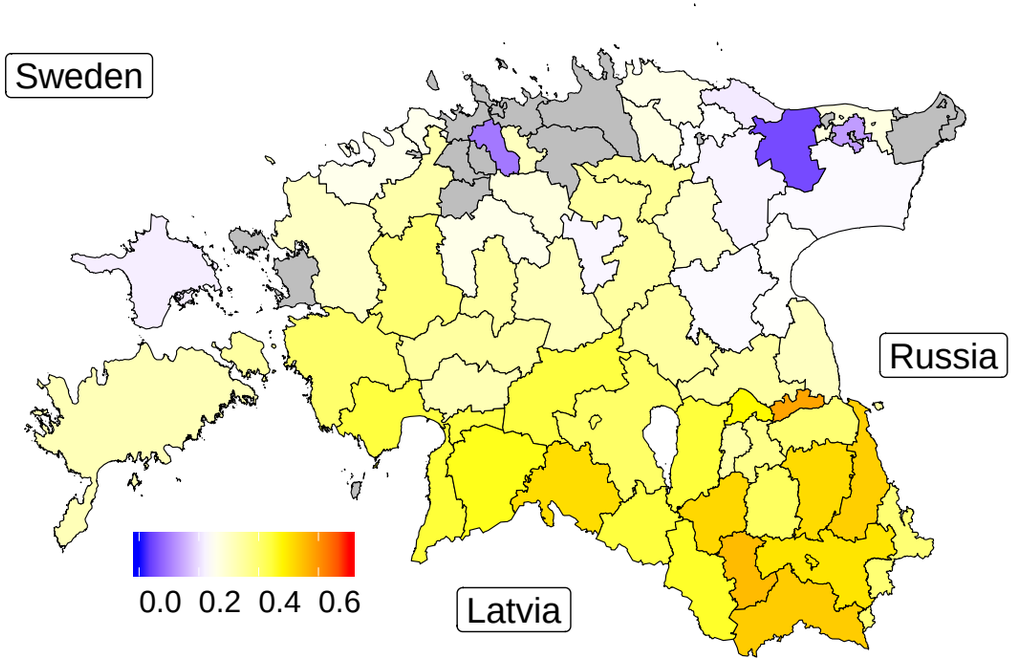
732Figure 6. Estonian effective population size dynamics. a, Effective population size estimates
733obtained by applying IBDNe19 to the entire dataset and to 4 clusters from Figure 5: eNW_1,
734eNE, eSW_2 and eSE_5. b, Comparison of historical and genetic estimates of Estonian
735population size. Historical estimates combine census data and reconstructions based on
736written or archaeological sources (S1:4.2-II). Genetic estimates are derived from IBDNe
737results, for which Est1527 subset was used (SI1:4.4-II) and refer to the broader population
738that contributed over time to the genomes of contemporary Estonians. When converting time
739points of the IBDNe curve into actual years we used the same logic as in the original
740publication19 and set generation 0 to correspond to the year when individuals in our sample
741had a mean age of 25 (1988). Generation time of 29 years was assumed. For year 1200 the
742minimum and maximum estimates are provided. In panel a shaded areas show 95%
743confidence intervals. In panel b shaded area corresponds to the range between the
744minimum and maximum genetic estimates of Nc (Methods), while the light blue line shows
745the geometric mean between the two. In both panels on the y axis, “k” stands for
746“thousands” and “M” for millions.

747Figure 7. Genome-wide plots of positive standardized SDS scores for the entire dataset (a)
748as well as SE (b) and nonSE (c) subsets. Conditional suggestive (blue) and genome-wide
749(red) significance lines are drawn. Gene names are highlighted for intragenic variants with –
750log10(p) > 5. Datasets are described in the text and Supplementary Information SI1:5.1.

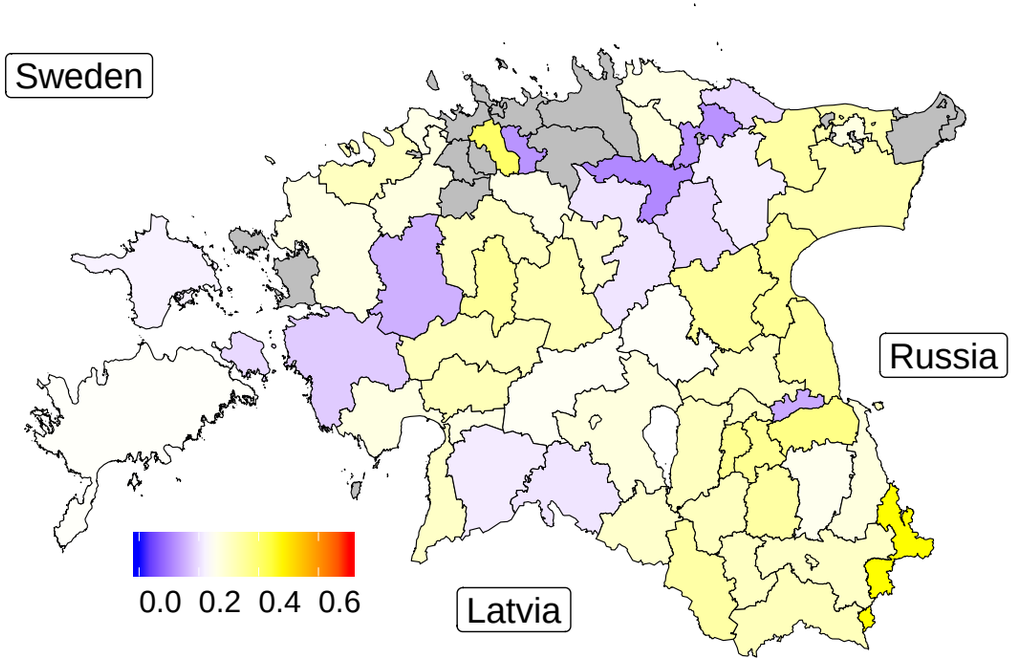

a
0.06
0.03
0.00
−0.03
−0.06
PC1, 0.14%
−0.05
0.00
0.05
0.10
PC2, 0.07%
SE
SW
NW
NE
noData
b
0.06
0.03
0.00
−0.03
PC2, 0.4%
−0.050
−0.025
0.000
0.025
PC1, 0.7%
Fin
Kar
Vep
Spa
Ita
Eng
Fre
Ger
Swe
Hun
RuN
RuCS
Mor
Ukr
Pol
Bel
Lit
Lat
0.025
0.000
−0.025
−0.050


a
b
E_1
E_2
NW_1
NW_2
NW_3
NW_4
NE_1
NE_2
SW_1
SW_2
SW_3
SW_4
SW_5
SW_6
SW_7
SE_1
SE_2
SE_3
SE_4
SE_5
SE_6
SE_7
31
11
9
7
67
18
12
52
7
14
9
15
14
20
12
18
29
43
28
5
17
30
10
20
30
40
50
60
70
80
cM

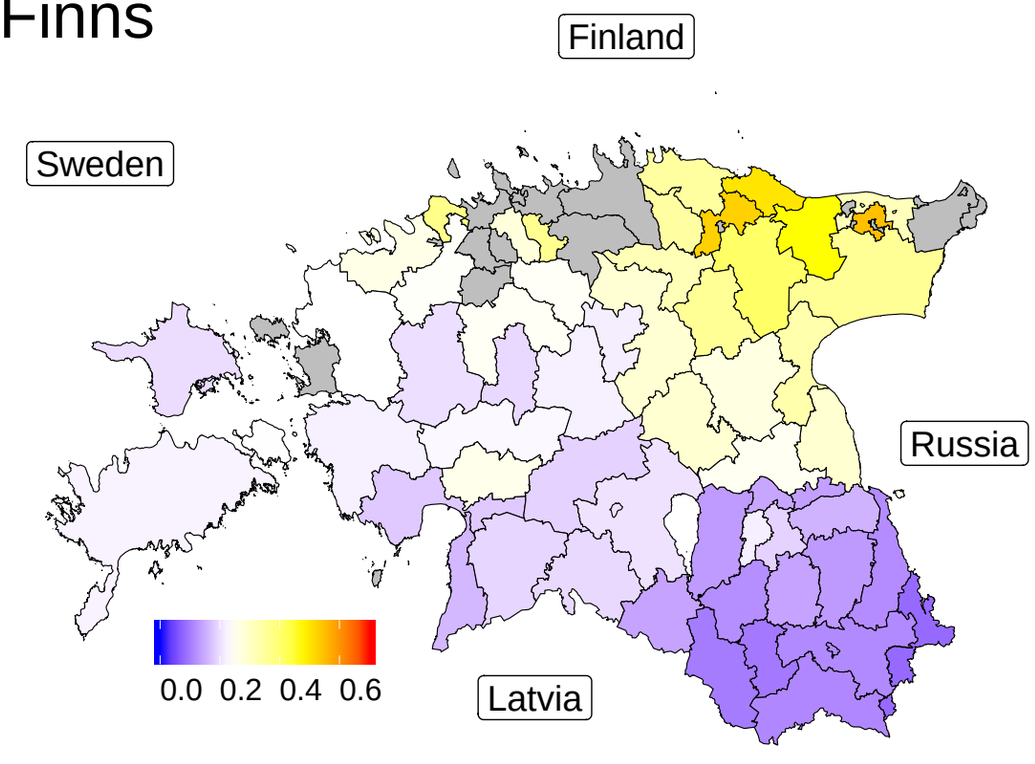
Balts



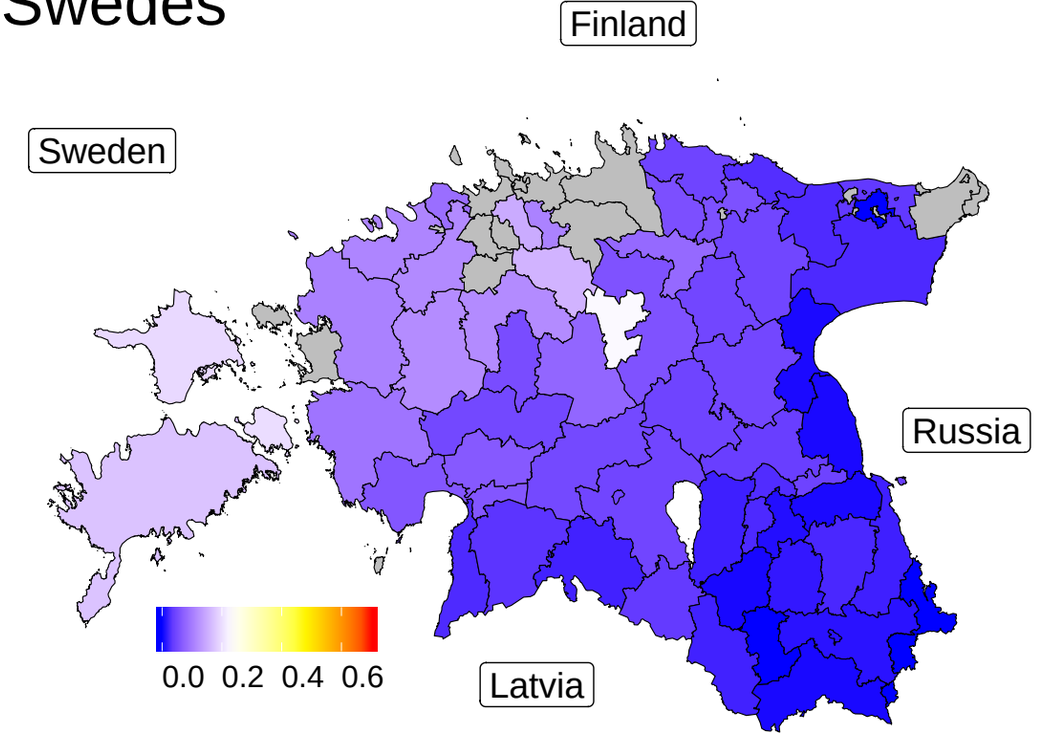
Slavs

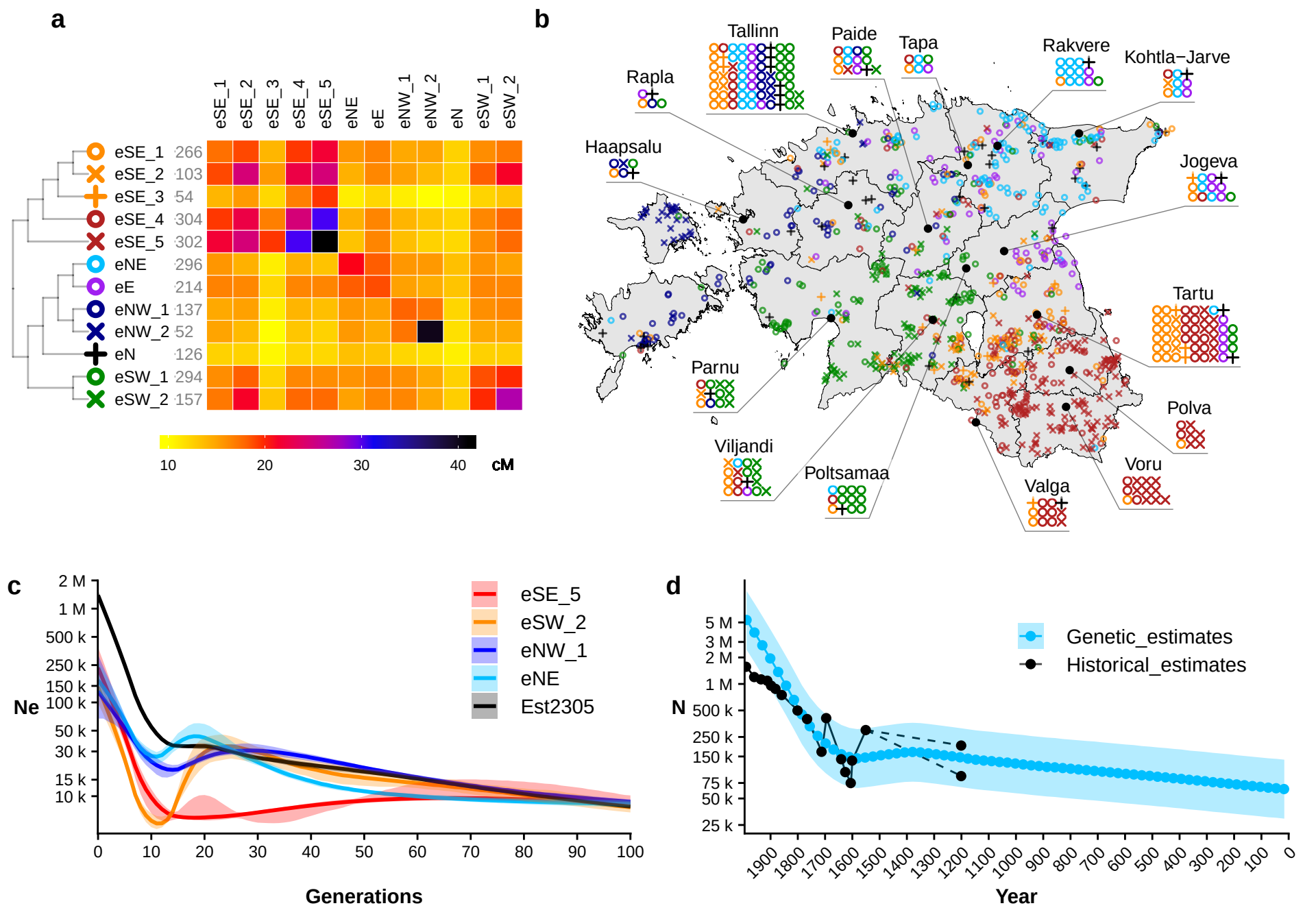


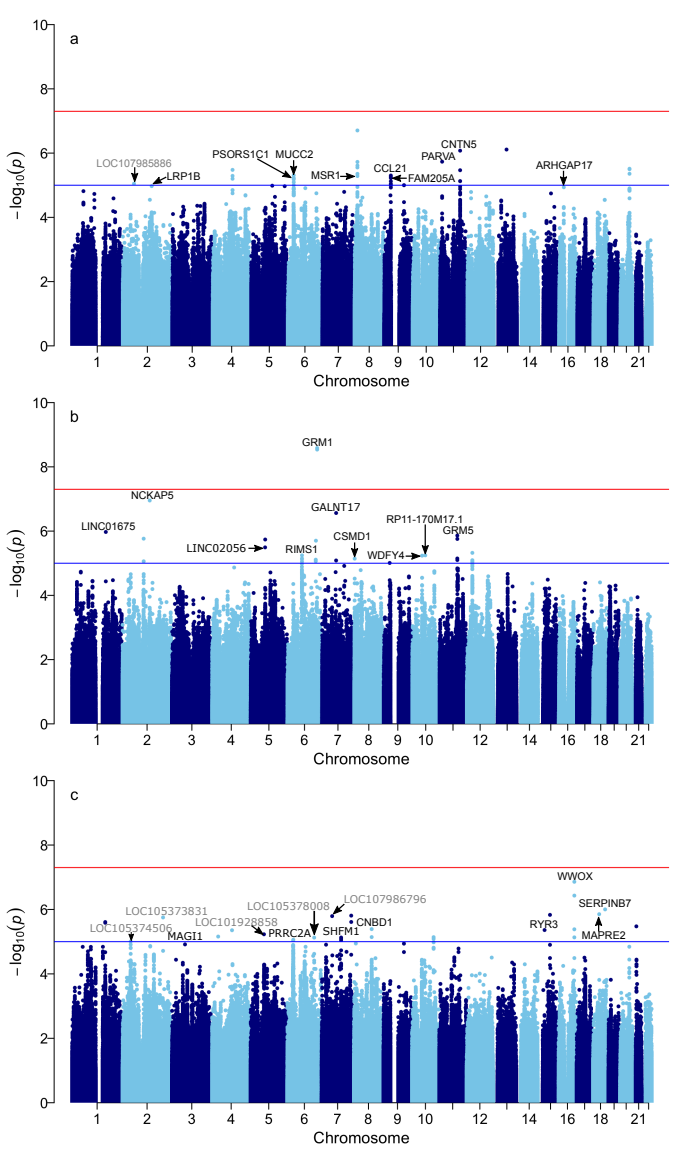
Finns



Swedes




a
eSE_1
eSE_2
eSE_3
eSE_4
eSE_5
eNE
eE
eNW_1
eNW_2
eN
eSW_1
eSW_2
266
103
54
304
302
296
214
137
52
126
294
157
10
20
30
40
cM
b
Tallinn
Paide
Tapa
Rakvere
Kohtla−Jarve
Rapla
Haapsalu
Jogeva
Tartu
Parnu
Polva
Viljandi
Poltsamaa
Voru
Valga
c
Ne
2 M
1 M
500 k
250 k
150 k
100 k
50 k
30 k
15 k
10 k
0
10
20
30
40
50
60
70
80
90
100
Generations
eSE_5
eSW_2
eNW_1
eNE
Est2305
d
N
5 M
3 M
2 M
1 M
500 k
250 k
150 k
75 k
50 k
25 k
1900
1800
1700
1600
1500
1400
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
Year
Genetic_estimates
Historical_estimates


a
b
c
$-\log_{10}(p)$
Chromosome
LOC107985886
LRP1B
PSORS1C1
MUCC2
MSR1
CCL21
FAM205A
PARVA
CNTN5
ARHGAP17
GRM1
NCKAP5
LINC01675
LINC02056
RIMS1
GALNT17
CSMD1
WDFY4
RP11-170M17.1
GRM5
LOC105373831
LOC105374506
MAGI1
LOC101928858
PRRC2A
LOC105378008
SHFM1
LOC107986796
CNBD1
WWOX
RYR3
SERPINB7
MAPRE2